House Price Prediction

▼ Data

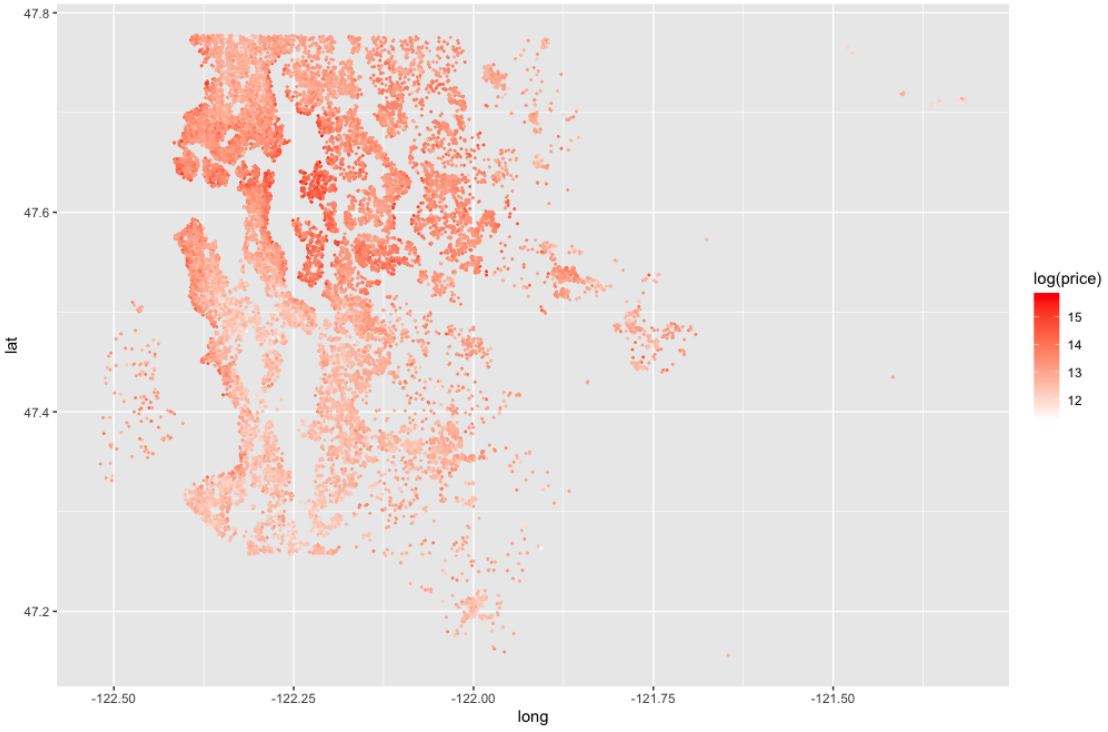
```
> head(house)
          id            date   price bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition
1 7129300520 20141013T000000  221900        3      1.00        1180     5650      1          0    0         3
2 6414100192 20141209T000000  538000        3      2.25        2570     7242      2          0    0         3
3 5631500400 20150225T000000  180000        2      1.00         770    10000      1          0    0         3
4 2487200875 20141209T000000  604000        4      3.00        1960     5000      1          0    0         5
5 1954400510 20150218T000000  510000        3      2.00        1680     8080      1          0    0         3
6 7237550310 20140512T000000 1225000        4      4.50        5420   101930      1          0    0         3
  grade sqft_above sqft_basement yr_built yr_renovated zipcode     lat     long sqft_living15 sqft_lot15
1     7       1180             0     1955            0   98178 47.5112 -122.257          1340       5650
2     7       2170           400     1951         1991   98125 47.7210 -122.319          1690       7639
3     6        770             0     1933            0   98028 47.7379 -122.233          2720       8062
4     7       1050           910     1965            0   98136 47.5208 -122.393          1360       5000
5     8       1680             0     1987            0   98074 47.6168 -122.045          1800       7503
6    11       3890          1530     2001            0   98053 47.6561 -122.005          4760     101930
```

▼ EDA

▼ EDA 1. 집의 위치와 가격의 관계 알아보기

📌 **Hypothesis 1** **집의 위치가 집의 가격에 영향을 미칠 것이다.**

👀 위도, 경도에 따른 집 값의 분포를 시각화



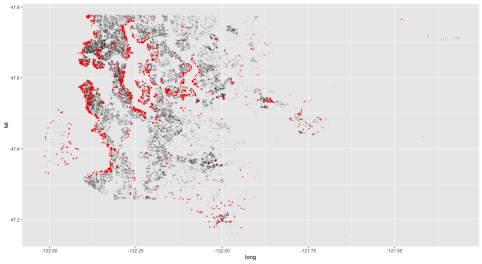
⭐ **Idea** 근처에 있는 집들끼리는 가격이 비슷하다 👉 어떤 집과 가깝다면 그 집의 가격과는 상관관계가 크고, 멀다면 상관관계가 작게 모델링

> ⭐ **Idea** 강을 근처에서 주변의 다른 집들보다 높은 가격대 형성 👉 강 근처인지 여부를 구별해줄 수 있는 변수만들기

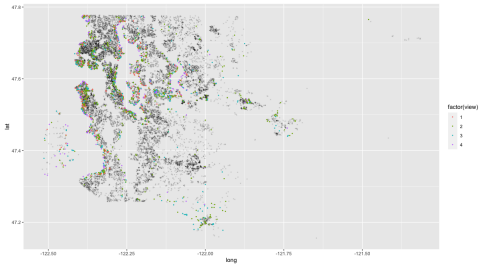
▼ view 변수를 이용해서 주변에 비해 높은 값을 가지는 집 찾기

view	0	1	2	3	4
count	19489	332	963	510	319

- view가 0이 아니라면 모두 동일하게 빨간색으로 시각화



- view의 그룹별로 다른 색깔을 보이도록 시각화



- view가 0이 아닌 값을 가진다면 근처의 view가 0인 집과 비교했을 때, 상대적으로 높은 가격대를 형성한다는 것을 알 수 있다.

▼ EDA 2. 집의 면적과 가격의 관계 알아보기

📌 **Hypothesis 2** **집의 면적이 집의 가격에 영향을 미칠 것이다.**

👀 집의 면적에 따른 집 값의 분포를 시각화

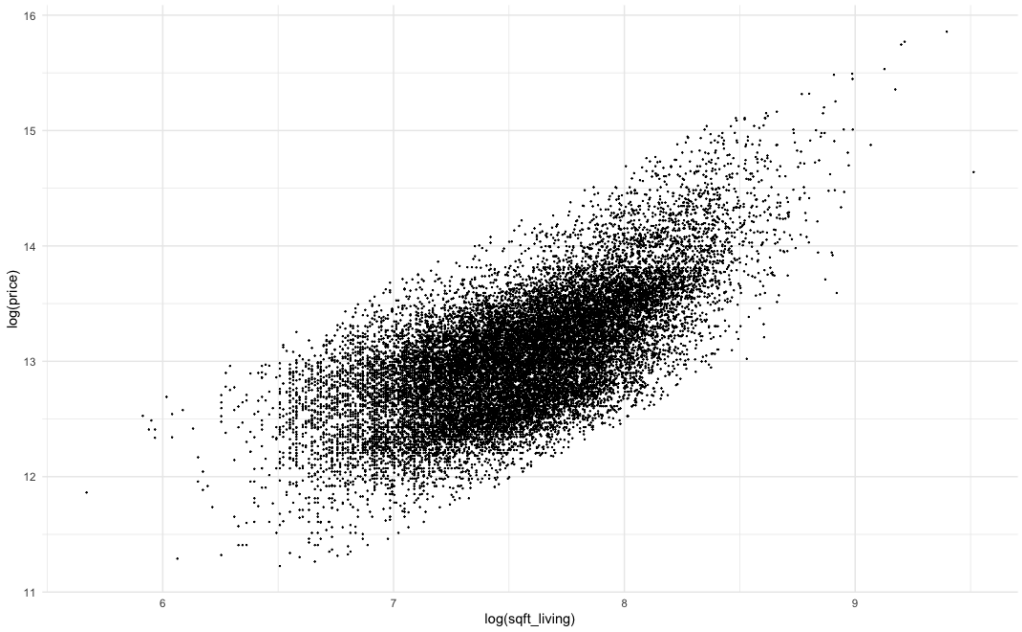


Figure 2. log(sqft_living) vs. log(price)

💡 **Insight 4** 면적이 증가할수록 가격도 증가하는 경향을 보이지만 같은 면적에서의 가격의 분산이 매우 크다 ⇒ 비슷한 면적일때의 가격의 분산을 설명해줄 다른 변수가 필요

👉 집의 면적과 함께 집의 quality를 반영할 수 있는 변수
grade, **view**, **condition** 를 사용하여 집의 면적에 따른 집 값의 분포를 다시 시각화

👉 EDA 1 에서 집의 가격에 영향을 주는것으로 보이는
lat 값을 고려했을 때의 집의 면적에 따른 집 값의 분포를 다시 시각화

▼ 👀 **grade 그룹별** 집의 면적과 집 값의 분포 시각화

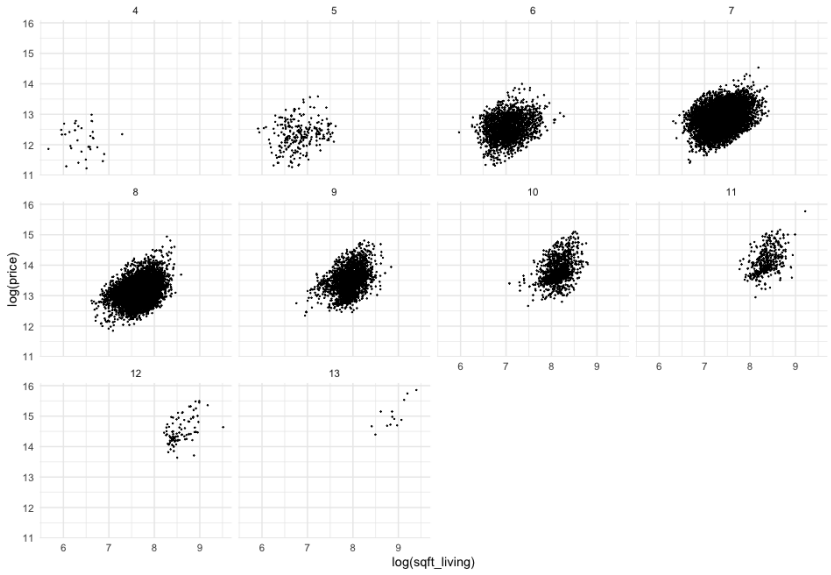
- grade 그룹별 count table

grade	1	3	4	5	6	7
count	1	3	29	242	2038	8981

  - grade가 1 또는 3인 집의 개수가 5개 이하로 매우 적기 때문에 grade가 1, 3, 4인 집을 합쳐 다시 그룹화한 변수 **grade2**를 만든다

- grade2 별 count table

grade	4	5	6	7	8	9
count	33	242	2038	8981	6068	2615



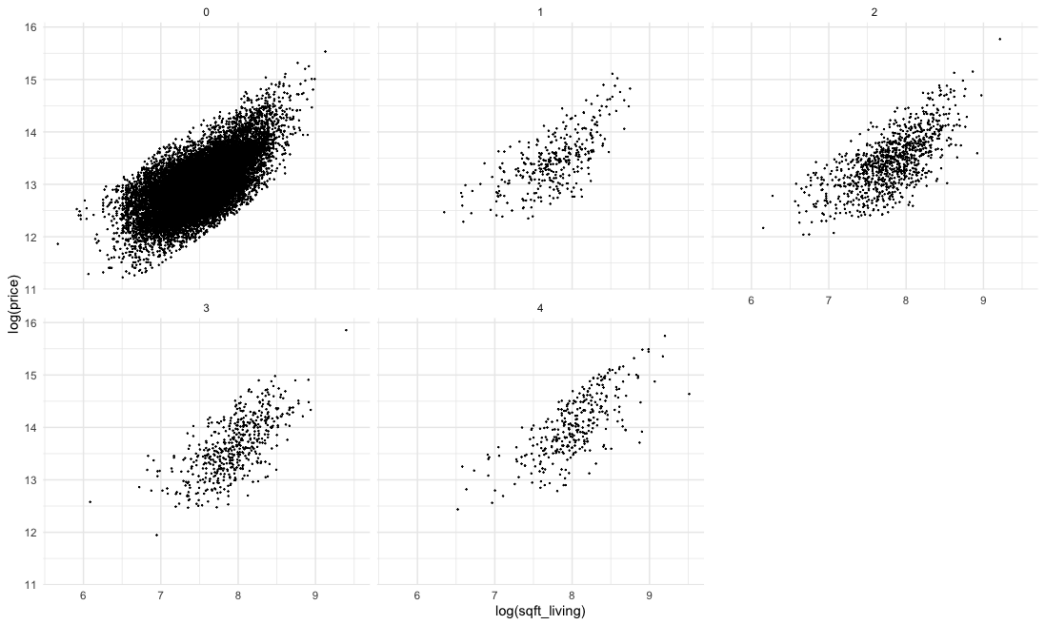
Scatter plot for each grade2 group

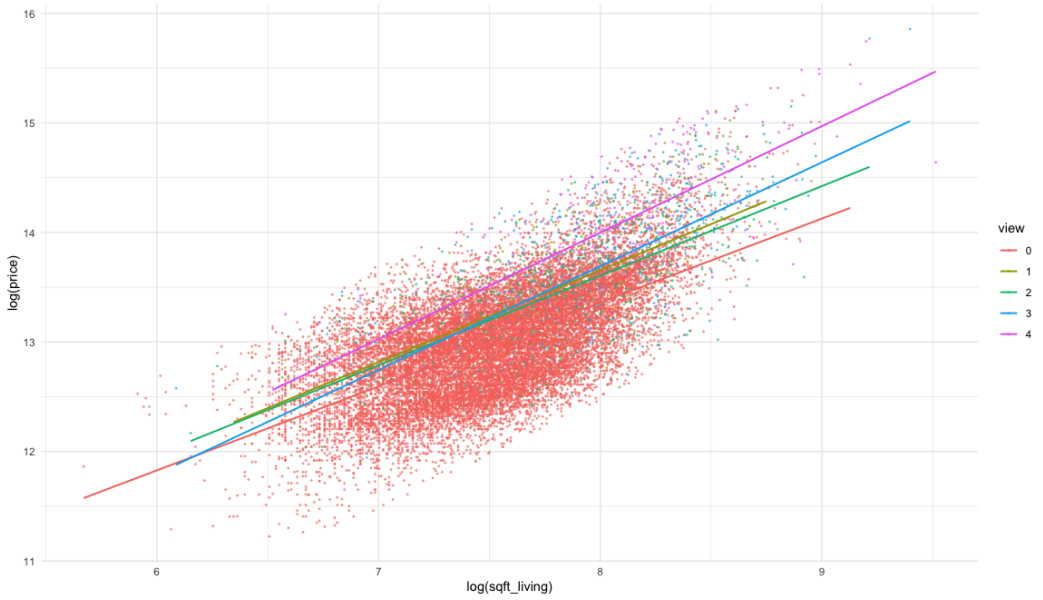
✔ **Conclusion** 그룹별로 봐도 여전히 point들이 뭉쳐있기 때문에 grade2 변수는 같은 면적 내에서의 가격의 변동성을 설명해주지 못한다.

▼ 👀 **view 그룹별** 집의 면적과 집 값의 분포 시각화

- view 그룹별 count table & scatter plot

view	0	1	2	3	4
#	19489	332	963	510	319

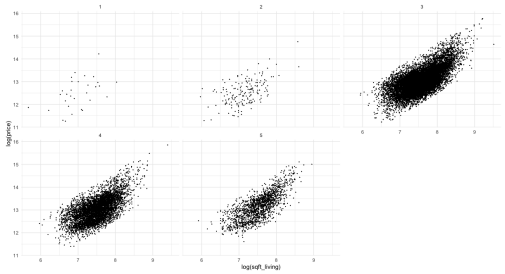


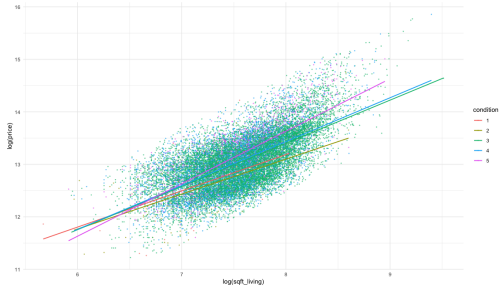


▼ 👀 **condition 그룹별** 집의 면적과 집 값의 분포 시각화

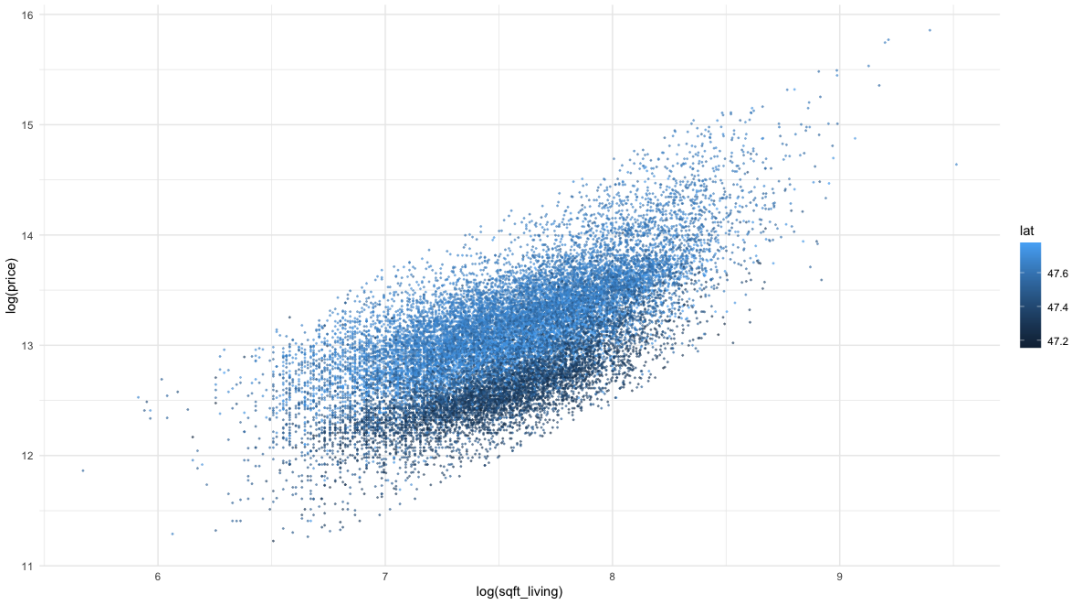
- condition 그룹별 count table & scatter plot & regression plot

condition	1	2	3	4	5
#	30	172	14031	5679	1701





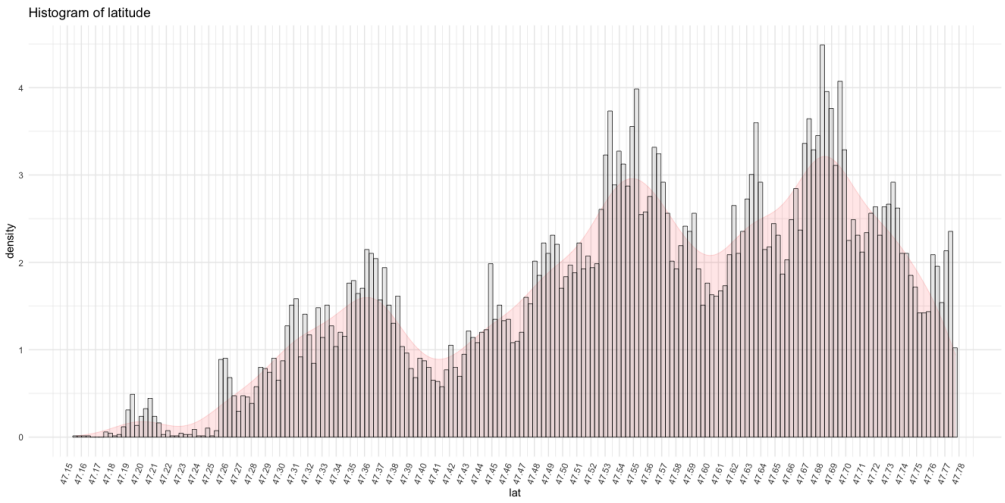
▼ 👀 **lat** 값을 고려한 집의 면적과 집 값의 분포 시각화



💡 **Insight** point의 색깔이 위아래로 구분되어 보임 👉 차이를 좀 더 구분해서 보기위해 lat 변수를 그룹화해서 다시 시각화

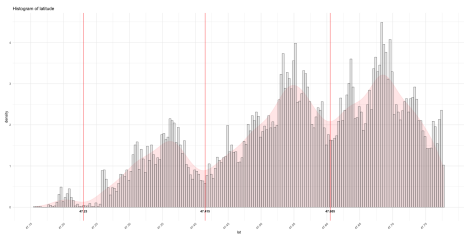
▼ 👀 lat 을 그룹화한 변수 **lat_group** 을 생성한후 lat_group 그룹별 집의 면적과 집 값의 분포 시각화

▼ 위도의 Histogram을 참고해서 그룹화



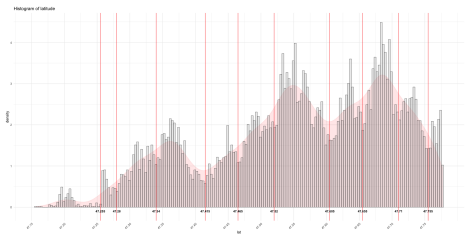
- lat_group 변수 만들기

  📌 **lat_group1** : 4개의 그룹으로 그룹화

  - 위도에 따른 집의 분포가 **정규분포 여러개가 혼합된 형태**를 보이므로 각 분포가 겹친다고 생각되는 부분을 기준으로 4개의 그룹으로 그룹화했다.

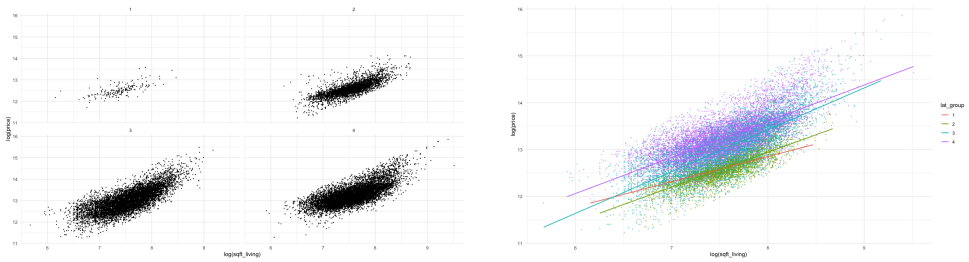
  📌 **lat_group2** : 9개의 그룹으로 그룹화

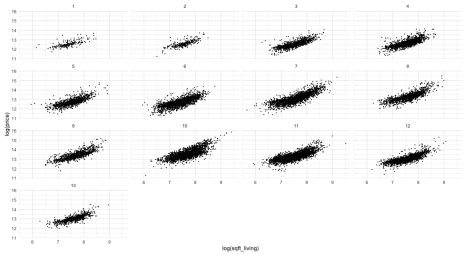
  - 더 많은 정규분포가 혼합되었다고 가정하고 더 세분화된 9개의 그룹으로 그룹화했다.

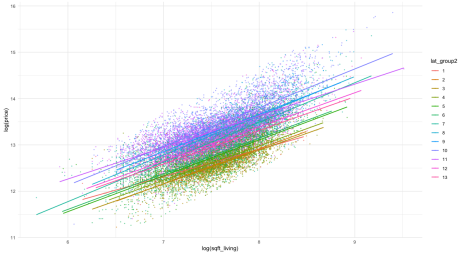
▼ lat_group 그룹별 scatter plot & regression plot

- lat_group1 그룹별



- lat_group2 그룹별





> **Conclusion** 각 그룹별로 기울기의 차이는 많이 없고 intercept 간의 차이가 더 뚜렷해보인다. 그룹별로 나누지 않았던 figure2 에 비해 더 직선에 가까운 형태를 보인다.

▼ Linear Regression 결과를 이용해 lat_group의 유의성 확인

- 전체 그룹의 회귀 직선에 대한 regression summary

$$y = m * \log(\text{sqft\_living}) + b$$

	Estimate	Std. Error	t value	p value	significance
b	6.7300	0.0471	143.0	<2e-16	***
m	0.8368	0.0062	134.5	<2e-16	***
Multiple R^2	0.4555	Adjusted R^2	0.4555		

- lat1_gruop 별 회귀 직선에 대한 regression summary

$$y = m * \log(\text{sqft\_living}) + m_2 * I_{(\text{lat1}=2)} + m_3 * I_{(\text{lat1}=3)} + m_4 * I_{(\text{lat1}=4)} + b$$

	Estimate	Std. Error	t value	p value	significance
b	6.4646	0.0432	149.4		
m	0.8198	0.0050	164.4		
m_2	-0.0425	0.0232	-1.834		
m_3	0.3681	0.0229	16.077		
m_4	0.6043	0.0229	26.42		
Multiple R^2	0.6513	Adjusted R^2	0.651		

- lat2_group 별 회귀 직선에 대한 regression summary

$$y = m * \log(\text{sqft\_living}) + \sum_{i=2}^{13} m_i * I_{(\text{lat2}=(i}$$

	Estimate		Std. Error
b	6.8273	***	0.0364
m	0.7733	.	0.0042
m_2	-0.0896	***	0.0237
m_3	-0.1060	***	0.0196
m_4	-0.0199		0.0192
m_5	0.0934		0.0198
m_6	0.1349		0.0192
m_7	0.4364		0.0191
m_8	0.5859		0.0197
m_9	0.6370		0.0200
m_{10}	0.7684		0.0191
m_{11}	0.6604		0.0188
m_{12}	0.3923		0.0192
m_{13}	0.3191		0.0202
Multiple R^2	0.7543	Adjusted R^2	

▼ lat_group에 따른 random intercept model 적합

```
> summary(model1)
Linear mixed model fit by REML ['lmerMod']
Formula: log(price) ~ log(sqft_living) + (1 + log(sqft_living) | lat_group1)
   Data: house

REML criterion at convergence: 10729.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.8515 -0.6433 -0.0479  0.5806  4.4055

Random effects:
 Groups     Name             Variance Std.Dev. Corr
 lat_group1 (Intercept)      0.73680  0.8584
            log(sqft_living) 0.01792  0.1339   -0.95
 Residual                    0.09596  0.3098
Number of obs: 21613, groups:  lat_group1, 4

Fixed effects:
                 Estimate Std. Error t value
(Intercept)       7.24995    0.44330   16.35
log(sqft_living)  0.74523    0.06859   10.87

Correlation of Fixed Effects:
            (Intr)
lg(sqft_lv) -0.956
```

```
> summary(model2)
Linear mixed model fit by REML ['lmerMod']
Formula: log(price) ~ log(sqft_living) + (1 + log(sqft_living) | lat_group2)
   Data: house

REML criterion at convergence: 3218.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.9526 -0.5972 -0.0440  0.5347  5.1528

Random effects:
 Groups     Name             Variance Std.Dev. Corr
 lat_group2 (Intercept)      0.18463  0.42968
            log(sqft_living) 0.00406  0.06371  -0.78
 Residual                    0.06752  0.25985
Number of obs: 21613, groups:  lat_group2, 13

Fixed effects:
                 Estimate Std. Error t value
(Intercept)       7.21186    0.12727   56.67
log(sqft_living)  0.76079    0.01864   40.82

Correlation of Fixed Effects:
            (Intr)
lg(sqft_lv) -0.800
```

> ⭐ **Idea** 면적과 위도는 집 값에 함께 영향을 미침 👉 면적은 위치와 함께 고려해서 모형에 반영

▼ Modeling

강 근처인지 여부를 구별해줄 수 있는 변수만들기 : good_view

👉 강을 근처에서 주변의 다른 집들보다 높은 가격대를 형성하는 경향을 반영해줄 수 있음

Gaussian Process(GP) 의 Multivariate Nonparametric Regression을 이용하고, covariance function으로 Isotropic covariance function을 사용

👉 GP를 사용함으로써 거리가 가까울수록 집의 가격 사이에 높은 correlation를 갖게할 수 있음

👉 Isotropic covariance function을 사용함으로써 위도와 경도 각각의 차이가 아닌 거리의 차이를 이용하여 correlation을 handling

👉 경도와 위도를 이용해 거리를 계산하는 방법인 Haversine formula를 covariance function으로 사용하여 두 집사이의 실제 거리를 이용하여 correlation을 handling

log(sqft_living) 변수를 이용한 선형모형을 추가

👉 GP를 이용해 위치에 따른 집 값의 변화를 고려한 후, 면적에 관한 변수를 이용해 면적에 따른 집 값의 변화를 capture할 수있음

Model

- Variables setting

Variable	Description
$y_i \in \mathbb{R}$	i 번째 집의 가격 $(i = 1, \dots, n)$
$\mathbf{z}_i = (z_{i1}, z_{i2}) \in \mathbb{R}^2$	i 번째 집의 경도(z_{i1})와 위도(z_{i2})
$x_{1i} \in \{0, 1\}$	i 번째 집의 good_view
$x_{2i} \in \mathbb{R}$	i 번째 집의 log(sqft_living)

  - $y_i \in \mathbb{R}$: i 번째 집의 가격 $(i = 1, \dots, n)$
  - $\mathbf{z}_i = (z_{i1}, z_{i2}) \in \mathbb{R}^2$: i 번째 집의 경도(z_{i1})와 위도(z_{i2})
  - $x_{1i} \in \{0, 1\}$: i 번째 집의 good_view

- $x_{2i} \in \mathbb{R}$: i 번째 집의 log(sqft_living)

- Model

$$y_i = f(\mathbf{z}_i) + \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0,\ \sigma^2)$$

$$f(\mathbf{z}_i) \sim \mathrm{N}\left(\mu(\mathbf{z}_i),\ \sigma^2\right), \quad i = 1, \ldots, n_k,\ k = 1, \ldots, K$$

$$(\mu(\mathbf{z}_1), \ldots \mu(\mathbf{z}_n)) \sim \mathrm{GP}\big(\,(m(\mathbf{z}_1), \ldots, m(\mathbf{z}_n))\,,\ K(\mathbf{z}_1, \ldots, \mathbf{z}_n)\big)$$

$$\text{where } k(\mathbf{z}_i,\ \mathbf{z}_j) = \tau^2 \exp\left(-\frac{d(\mathbf{z}_i,\ \mathbf{z}_j)}{l^2}\right)$$

$$d(\mathbf{z}_i,\ \mathbf{z}_j) = R \cdot c_{ij},\ R = 6371.0,$$

$$a_{ij} = \sin^2\left(\frac{z_{i2} - z_{j2}}{2}\right) + \cos(z_{i2}) \cdot \cos(z_{j2}) \cdot \sin^2\left(\frac{z_{i1} - z_{j1}}{2}\right)$$

$$c_{ij} = 2 \cdot \mathrm{atan2}(\sqrt{a_{ij}}, \sqrt{1 - a_{ij}})$$

$$\mathrm{atan2}(y, x) = \begin{cases} \arctan\left(\frac{y}{x}\right) & \text{if } x > 0 \\ \arctan\left(\frac{y}{x}\right) + \pi & \text{if } x < 0, y \geq 0 \\ \arctan\left(\frac{y}{x}\right) - \pi & \text{if } x < 0, y < 0 \\ +\frac{\pi}{2} & \text{if } x = 0, y > 0 \\ -\frac{\pi}{2} & \text{if } x = 0, y < 0 \\ \text{undefined} & \text{if } x = 0, y = 0 \end{cases}$$

$$\sigma^2 \sim \text{inv-Gamma}(a_\sigma,\ b_\sigma),$$

$$p(\sigma^2) \propto 1,$$